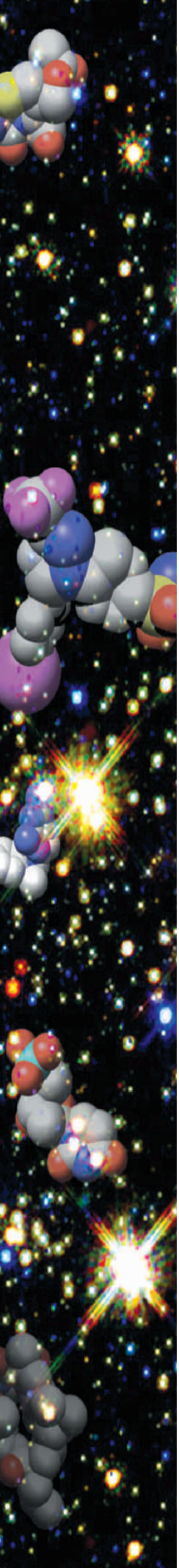

Chemical space and biology

Christopher M. Dobson

Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, UK (e-mail: cmd44@cam.ac.uk)

Chemical space — which encompasses all possible small organic molecules, including those present in biological systems — is vast. So vast, in fact, that so far only a tiny fraction of it has been explored. Nevertheless, these explorations have greatly enhanced our understanding of biology, and have led to the development of many of today's drugs. The discovery of new bioactive molecules, facilitated by a deeper understanding of the nature of the regions of chemical space that are relevant to biology, will advance our knowledge of biological processes and lead to new strategies to treat disease.

Living systems have evolved over several billion years to carry out carefully controlled chemistry in an aqueous environment at temperatures almost exclusively between zero and 100 °C. Under these conditions and unaided, many of the chemical reactions that are essential to life would not occur at perceptible rates, and most would not result in specific and reproducible products. Enzymes, along with other proteins and some nucleic acids, are used by natural biological systems to achieve this control; these macromolecules are responsible for the synthesis, transport and degradation of virtually every chemical compound in the biological environment[1].

However, the chemical compounds used by biological systems represent a staggeringly small fraction of the total possible number of small carbon-based compounds with molecular masses in the same range as those of living systems (that is, less than about 500 daltons). Some estimates of this number are in excess of 10^{60} (ref. 2). The simplest living organisms can function with just a few hundred different types of such molecule, and fewer than 100 account for nearly the entire molecular pool[3,4]. Moreover, it seems that the total number of different small molecules within our own bodies could be just a few thousand[4]. So, it is clear that, at least in terms of numbers of compounds, 'biologically relevant chemical space' is only a minute fraction of complete 'chemical space' (see Box 1 for a definition of the terms used in this Insight). It is remarkable that so many complex processes can be carried out with such a limited number of molecules, and that biological chemistry can be so rich and diverse despite the relatively limited range of reactions that seem to have been exploited during the evolution of living systems (see Box 2 for a discussion of why particular types of chemistry might have emerged as the basis of life).

Similarly, as revealed by the recent triumphs of a variety of international sequencing projects, the genomes of the simplest living systems encode the sequences of less than 1,000 different proteins and the human genome about 100 times more[5] — numbers that are minute when compared with the total number of proteins that could theoretically exist. As there are 20 different types of amino acid and the average size of a natural protein is about 300 residues, this number is a staggering 20^{300} or more than 10^{390}, and if only a single molecule of each of these polypeptides were to be produced, their combined mass would vastly exceed that of the known universe. Natural proteins are therefore also a very select group of molecules.

The characteristics of this select group of natural proteins are linked to those of the small molecules that are used in living systems, and to those of the relatively small number of synthetic small molecules that we have developed into drugs. Understanding this link will help us answer the question of how we can best use the powerful new methods that are emerging to probe biological systems, both to understand the fundamental processes of life and to develop new strategies to treat disease.

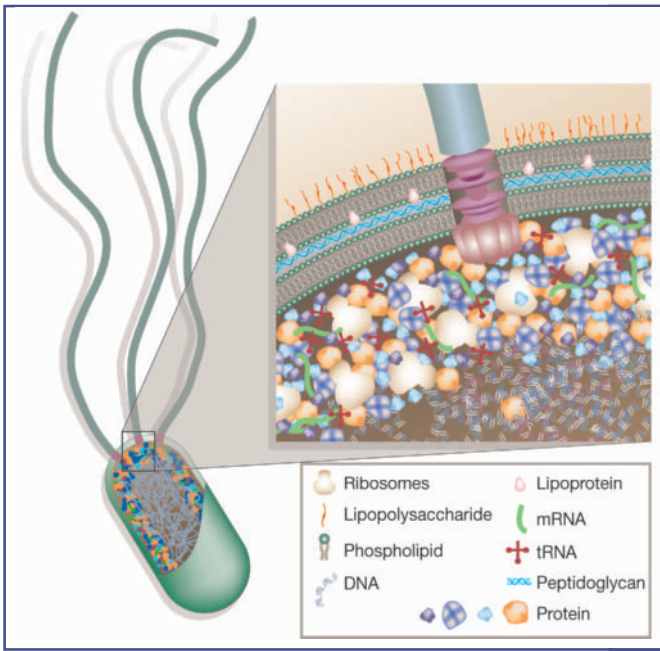

Figure 1 Schematic representation of a crowded cell. An array of different molecules can function independently under extremely crowded conditions, partly because of judicious distributions of oppositely charged polar groups on the molecular surfaces[38]. However, such systems are in some ways extremely fragile. For example, a mutation that alters just one amino acid in the haemoglobin molecule (replacing a charged carboxylic acid with a methyl group) can stimulate massive aggregation and give rise to a fatal genetic disease, sickle-cell anaemia[8,39]. More generally, many disorders of old age, most famously Alzheimer's disease, result from the increasingly facile conversion of normally soluble proteins into intractable deposits that occur particularly as we get older (see http://www.horizonsymposia.com/ for the Horizon Symposium 'Protein Folding and Disease', and ref. 40). Many of these aggregation processes involve the reversion of the unique biologically active forms of polypeptide chains into a generic and non-functional 'chemical' form[41]. Adapted with permission from D. Goodsell.

Box 1

Glossary of important terms relevant to chemical space and biology

Bioavailability
The fraction or percentage of an administered drug or other substance that becomes available to the target tissue after administration.

Biologically relevant chemical space
Those parts of chemical space in which biologically active compounds reside.

Chemical genetics
The study of gene-product function in a cellular or organismal context using a set of exogenous ligands, often known as chemical tools.

Chemical library
A collection of chemical compounds.

Chemical space
Chemicals can be characterized by a wide range of 'descriptors', such as their molecular mass, lipophilicity (their affinity for a lipid environment) and topological features. 'Chemical space' is a term often used in place of 'multi-dimensional descriptor space': it is a region defined by a particular choice of descriptors and the limits placed on them. In the context of this Insight, chemical space is defined as the total descriptor space that encompasses all the small carbon-based molecules that could in principle be created.

Combinatorial chemistry
The generation of large collections or 'libraries' of compounds by combinations of a set of smaller chemical structures, known as 'building blocks'.

Druggability/druggable target
The feasibility with which a macromolecular target can be modulated by a small molecule that has appropriate properties to be developed into a drug.

Drug-like
Sharing certain characteristics with other molecules that act as drugs. The exact set of characteristics — size, shape and solubility in water and organic solvents — varies depending on who is evaluating the molecules.

Genome
All the genetic material in the chromosomes of a particular organism.

High-throughput screening
In high-throughput screening, large libraries of chemical compounds (typically 10,000 to 100,000) are screened in a biological assay, for example, for their ability to bind to a particular protein or to inhibit a particular cellular process.

Hit
An active compound that exceeds a certain threshold value in a given assay; for example, more than 90% inhibition of an enzyme's activity.

Lead
A chemical structure or series of structures that demonstrate activity and selectivity in a biological screen. In drug discovery, a lead is used as a basis for chemical optimization, with the aim of identifying a clinical candidate.

Lipinski's rules
Lipinski's analysis of the World Drug Index led to the 'rule of five'[15]. This identifies several key properties that should be considered for small molecules that are intended to be orally administered. These properties are: molecular mass less than 500 daltons; number of hydrogen-bond donors less than 5; number of hydrogen-bond acceptors less than 10; calculated octanol/water partition coefficient (an indication of the ability of a molecule to cross biological membranes) less than 5.

Natural product
A chemical substance produced by a living organism. This term is often used in reference to small chemical substances found in nature that have distinct pharmacological effects, such as the antibiotic penicillin.

Proteome
The complete set of proteins that can be expressed by the genetic material of an organism.

RNA interference (RNAi)
A process by which double-stranded RNA silences specifically the expression of homologous genes.

Chemistry in a biological environment

A crucial factor in understanding the nature of living systems is that biological molecules do not act in isolation in the dilute solutions familiar to most chemists. Instead, they are packed together to an extraordinary degree within cells[6,7]. Indeed, the concentration of macromolecules inside cells can amount to several hundred grams per litre. Many of us may have been astonished during our school days to learn that our bodies are more than 70% water, but how many of us wondered at the difficulty of making a 30% solution of molecules that are rich in hydrocarbon derivatives and other hydrophobic groups? A space-filling representation of a typical cell (Fig. 1) illustrates how molecular species are crowded together in its complex organizational structure[8,9]. Such 'molecular crowding' is likely to be important in many facets of biological chemistry. For example, binding affinities and the rates of self-assembly can change by orders of magnitude as a result of this phenomenon. Crowding is therefore an important factor to consider when using data derived from *in vitro* studies in dilute solution to understand processes taking place *in vivo*[6,7]. Moreover, biological systems are increasingly being considered as highly interconnected sets of interactions (as shown, for example, by the emergence of 'systems biology') in contrast to the reductionist view of much of traditional biochemistry[10]. In addition, considerable efforts are being made to understand the astonishing ability of biological molecules to self-assemble and generate functional entities ranging from folded proteins to whole organisms[11].

Techniques such as X-ray crystallography, nuclear magnetic resonance (NMR) and mass spectrometry have already revolutionized our understanding of the structure and function of biological molecules. It is now becoming possible to examine the ultrastructure of cells in remarkable detail, primarily through the development of modern imaging techniques[12]. Of particular importance are methods based on fluorescence emission. These can be used together with confocal microscopy to identify and track an increasingly wide range of molecules (both large and small) within their biological environments. Perhaps the most dramatic technique, however, is that based on electron microscopy: 'cryoelectron tomography' is now beginning to allow us to visualize, within a cell, molecular assemblies such as actin, which provides cells with their internal structures, and ribosomes, the complexes of proteins and nucleic acids that are responsible for all protein synthesis[13]. Along with these experimental approaches, computational procedures are being developed to simulate the behaviour of molecules within whole cells or indeed whole organisms[14]. Further developments of this type will undoubtedly lead to a deeper understanding of how cellular components of all types interact with each other. Even without

such information, however, the high density of molecules in cells is a remarkable phenomenon that must be borne in mind when we attempt to perturb their behaviour for therapeutic purposes.

The challenges of drug discovery

Although some therapeutic agents are designed to increase the natural concentrations of key biological molecules that are depleted in particular disease states (for example, insulin), the primary objective of most pharmaceutical chemistry is to generate new compounds that can modulate disease processes. Most prized are relatively small molecules (only a small percentage of orally administered drugs have molecular masses above 500 daltons[15]) whose properties enable them to interact with and perturb the function of given biological molecules. It is equally important, however, that these compounds do not interact with most other molecules and generate potentially adverse side effects. The immensity of this task is illustrated by the schematic illustration in Fig. 1.

Box 2

The chemistry of life

One of the most fundamental questions relating to biological diversity is why particular types of molecule have emerged as those on which the chemistry of all life forms is based. It is clear that solubility in water is a key issue. Although 99% of the atoms within a biological system are C, H, O or N, more than 20 other elements are essential to life. All these elements are (or were when life on Earth began) relatively abundant in the Earth's crust, the sea or the atmosphere, and their ions or common compounds are soluble in water[42]. Solubility in water is also likely to be a major reason why many of the small organic molecules used by biological systems (including the amino acids) are derivatives of simple carboxylic acids and organic amines; these groups are normally charged, and therefore hydrophilic, at physiological pH. Similarly, many others are charged derivatives of phosphoric acid[43], the chemical entity that is also the precursor of ATP, the chief energy store in biology, and the scaffold for DNA and RNA. The unique properties of water also cause other derivatives of phosphoric acid, the phospholipids, to assemble into bilayers that are the key components of all biological membranes. The energetic advantage of burying hydrophobic groups away from water in the interior of a closely packed structure is also an important driving force in protein folding[1,41]. To allow folding, a significant proportion of the 20 amino-acid side chains incorporated into natural proteins are very hydrophobic, and the rest, many of which end up on the surface of folded proteins, are to varying degrees hydrophilic.

The chemical properties of the various side chains of proteins, along with a selection of metal ions and cofactors that can be incorporated into the folded structures, not only permit folding but also define the fundamental chemistry of life. The side chains of the natural amino acids, which are the same in every living organism, contain only a small selection of the functional groups that are familiar from any chemistry textbook: a methyl (but not an ethyl) group; an isopropyl (but not an *n*-propyl) group; a primary and a secondary alcohol; a thiol and an imidazole group; two carboxylic acids and so on[39]. But why this particular set of 20 chemical groups? Do these groups have the unique range of properties required to catalyse all the reactions needed for life to occur? Or did they arise by chance and has life on Earth been too short to allow the evolution of a wider range of chemical entities? The answers to such questions have long been the subject of speculation, but are now beginning to be probed directly by experiment. One remarkable new approach exploits the usual mechanism of protein synthesis in bacteria to generate proteins containing new types of amino acid[44]. It will be fascinating to learn what additional chemical tasks such organisms can perform, and how they respond to selective pressure in laboratory experiments that simulate natural evolution. Undoubtedly, such forays into 'abnormal' biology will shed light on 'normal' biological evolution and function, and indeed on the types of novel chemical entity that can interact selectively with natural biomolecules.

The natural products of different organisms — largely plants and bacteria — or their derivatives have been the staple tools of healers from the dawn of history until the birth of modern synthetic chemistry in the nineteenth century. Now, with the immense developments in combinatorial methods over the past decade or so, huge arrays of new molecules can be produced in relatively short periods of time[16,17]. Together with rapid screening methods, the drug-discovery process has been moving into uncharted territory; seemingly endless numbers of potentially active compounds are becoming available. As our knowledge of even the most complex aspects of biology at a molecular level expands, we can increasingly use rational arguments in the design of potential therapies and of new molecules that are promising to test or screen[18]. Despite such expert knowledge, the scale of the procedures needed to find appropriate compounds is remarkable; some individual drug companies screen millions of potential compounds each year against a range of targets, and even then, success is not guaranteed. As we have seen, however, such numbers are insignificant compared with the total number of possible small organic molecules. In addition, even the biggest libraries of compounds used in screening may not reflect the rich chemical diversity of the much smaller numbers of natural products[19] (Fig. 2). It is clear, therefore, that reliable computational approaches to sift through much larger numbers of more varied compounds would be of tremendous value in drug discovery. Once likely candidates for a given purpose are identified, experimental screening procedures could then be focused on a much smaller range of selected compounds. As Shoichet discusses in a commentary in this issue (page 862), the examination of molecules *in silico* for their ability to bind to specific targets already plays an important part in screening strategies, although such 'virtual screening' approaches have yet to achieve their full potential in the drug-discovery process.

Despite the many advances in technology, the cost of generating new drugs is inexorably rising, leading to ever greater pressure on pharmaceutical companies to focus on developing therapies primarily for the common diseases of wealthy countries[20,21]. Those suffering from rare diseases, and indeed the vast number of people in poorer countries, particularly in the tropics, are all too often neglected in the continuing fight against infection and disease. But despite the evidence that the new techniques entering the pharmaceutical industry have not yet been a panacea for the drug-discovery process[22], it is still early days. We have yet, for example, to reap the real benefits of the recent revolutions in genomics and proteomics, which promise to identify a much greater number of well-characterized molecular targets for therapeutic intervention[23]. Indeed, the number of new targets that have emerged in recent years within the pharmaceutical industry as a whole is remarkably small. For example, between 1994 and 2001, just 22 drugs that modulate new targets were approved[24]. So far, analyses have revealed that the total number of human proteins against which drugs have been targeted is less than 500 (ref. 25), a small percentage of the estimated total number of proteins in the human body. Although expert opinions differ as to the total number of possible 'druggable' targets, it is certainly larger than the number currently known[25,26].

Chemical 'tools' for biological systems

One of the potential problems with the new types of organic compound that are now being explored as drugs is that they may be extremely potent when tested against isolated targets in the laboratory environment, but within the complex cellular milieu (Fig. 1), they might interact with cellular components other than the desired target. The small molecules found naturally in biological systems, often called 'natural products', have at least been through the evolutionary mill and are perhaps less likely to interact in a damaging manner with common components of living systems, such as membranes or DNA. Indeed, of all drugs licensed over the past 20 years, around 30% are natural products or natural-product derivatives. If we include compounds 'inspired by' natural products, the fraction rises to almost twice this number[27] (see also the review in this

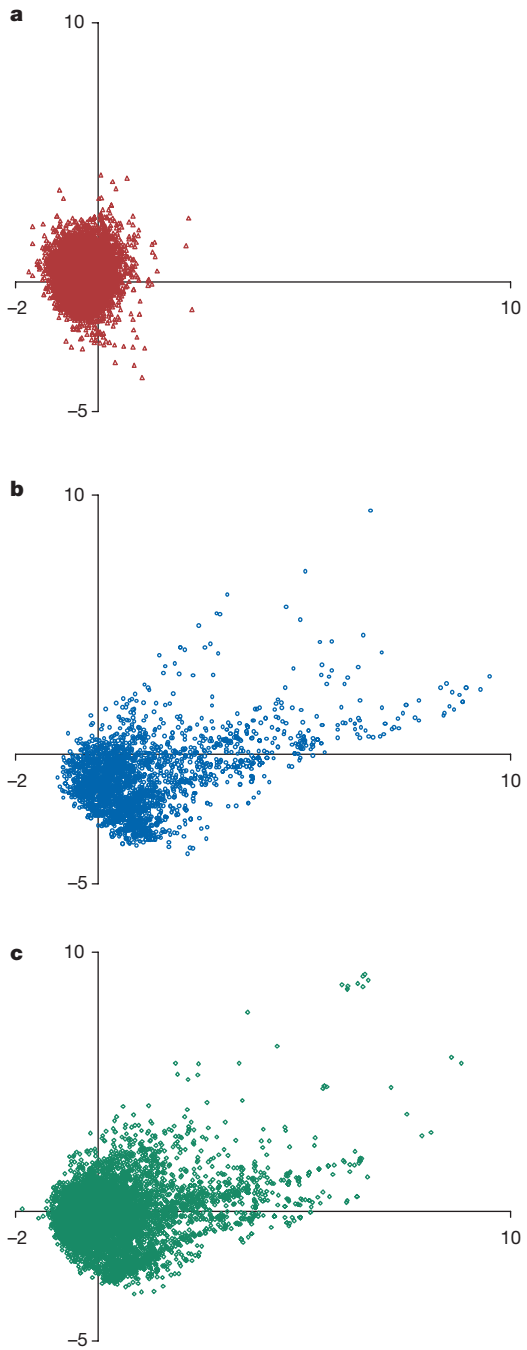

Figure 2 Comparison of the properties of different classes of molecule. A large database that contained compounds from combinatorial chemistry (**a**), natural products (**b**) and drugs (**c**) was analysed on the basis of a variety of molecular properties[19]. To visualize the diversity of these compounds on the basis of these properties, a statistical approach known as principal component analysis was used. Plots of the first two principal components — which explain about 54% of the variance in the properties analysed — are shown. Combinatorial compounds cover a well-defined region in diversity space given by these principal components. Both drugs and natural products cover all this space, as well as a much larger additional region of space. It is of particular interest to note the similarity of the plots of natural products and successful drug molecules. Adapted with permission from ref. 19.

issue by Clardy and Walsh, page 829). Interestingly, a comparison of the properties of drugs, natural products and combinatorial chemistry libraries shows that combinatorial compounds typically cover a significantly smaller area of chemical space than either drugs or natural products[19] (Fig. 2). This suggests that by aiming to mimic some properties of natural compounds, new combinatorial compounds could be made that are substantially more diverse and that have greater biological relevance[19] than those currently known.

Remarkably, however, it has been estimated that only 0.1% of all bacterial strains — the richest source of new biological molecules — has been cultured and analysed[28]. Thus, as Clardy and Walsh discuss in this issue (page 829), there is a vast harvest of new natural products, perhaps running to millions of new compounds, waiting to be gathered from previously unexplored strains of living organisms (mainly bacteria, plants and fungi). Moreover, there are now opportunities to manipulate nature's 'production lines', for example, by using mutagenesis and gene shuffling to induce microorganisms to create new biologically active molecules, and hence to generate large libraries of new 'natural products'.

One of the most important aspects of the development of new techniques and technologies is that they can be used for two distinct but highly complementary purposes. The focus of most activity in academic environments is to use these new approaches to understand the fundamental basis of cellular and organismal biology. The primary objective of most industrial research, however, is to use such strategies to discover new drugs, or at least new lead compounds for drug discovery. These activities are not of course mutually exclusive, and indeed closer interactions between members of these two communities could bring substantial benefits to both parties.

The use of the vast libraries of new small molecules as 'chemical tools' to probe biological function and discover potential therapeutics is discussed in the reviews in this issue by Stockwell (page 846), and Lipinski and Hopkins (page 855). Using small molecules to probe biological systems is now often described as 'chemical genetics' or 'chemical genomics'[29]. The enormous complexity of the biological milieu, again evident in Fig. 1, makes one of the ultimate goals of this approach — to discover a small molecule to modulate the function of every protein — an extremely challenging task, even in the light of the large arrays of chemical compounds that can be generated by combinatorial methods of ever-increasing sophistication. As well as the issues of diversity and specificity, cells may have evolved mechanisms to protect some of their most vital proteins from interference by small, extraneous molecules. Another major issue in chemical genetics concerns the quality of the data that are generated using various assay technologies; screening the same biological target with three different types of assay was recently found to give a set of hits that is consistent from assay to assay in only about 30% of cases[30]. Although such a low level of consistency may not be very important for drug discovery, where the main objective is often simply to identify a number of active compounds, it can be debilitating if the objective is to chart the network of interactions within a biological organism. The quality of the chemical libraries and the reliability of screening techniques are still limiting factors in our knowledge of biological systems and their molecular diversity.

In addition to using the products of synthetic organic chemistry as tools to probe biological systems, new molecular tools based on other cellular components, such as DNA and RNA, are increasingly being developed. As Breaker discusses in a review in this issue (page 838), various RNA technologies are currently generating a great deal of interest. That RNA molecules play an important part in biological chemistry is well established, notably as the catalytic ribozymes that are involved in many important biological reactions, not least protein synthesis[31]. Moreover, RNA interference (RNAi), in which synthetic RNA fragments are designed to interfere with the normal expression of specific genes, is becoming an important tool for exploring gene function, as discussed at a recent Horizon Symposium, 'Understanding the RNAissance' (http//:www.horizonsymposia.com), and reported

in ref. 32. In addition, aptamers — RNA molecules that form binding pockets for ligands with specificities and affinities similar to those of antibodies — are emerging as new probes of the functions of both large and small molecules. Aptamers that bind to particular targets can be engineered using *in vitro* evolution and amplification techniques. They can then be used as reagents to probe the roles of specific molecules in a given biological system. Furthermore, members of a previously neglected class of molecules, the oligosaccharides, are emerging as biological tools, now that efficient methods for sequencing and synthesizing these complex molecules are being developed[33]. In addition to acting as probes of biological function and regulation, all these types of molecule are themselves becoming the focus of drug discovery efforts.

Future prospects

A rich array of data on the effects of small molecules on biological systems is accumulating, mainly from large-scale screening exercises (although the quality of this information is often less than optimal; see the review in this issue by Lipinski and Hopkins, page 855). Analysis of such databases, using the types of computational method pioneered by the flourishing bioinformatics community[34], should lead to major advances, both in our understanding of biological chemistry and in our ability to identify promising therapeutic compounds and therapeutic targets[35]. Although progress is now being made in developing tools for mining chemical information, such progress is often limited by the difficulty in accessing much of the data of interest[36]. Some estimates suggest that only about 1% of some types of chemical information are in the public domain. In contrast, the majority of many forms of biological data, from gene sequences to protein structures, is freely accessible to scientists in both academia and industry. One of the reasons for the inaccessibility of so much chemical information, in addition to the technical challenges of cataloguing and checking vast amounts of data, is concerned with issues of intellectual property. However, one can be optimistic that ways will be found to overcome the various hurdles to allow these resources to be used in the most effective ways possible.

With increasingly diverse, reliable and accessible databases of information about the effects of new chemical compounds on specific biochemical processes, we shall be able to understand much more about the nature of biologically relevant chemical space. In addition, we shall learn more about the types of compound that might make good drugs by analysing the behaviour of a much wider range of small molecules than the miserly number used by our bodies for so many purposes — from generating energy to building arsenals of macromolecules. In this regard, among the most exciting recent developments are efforts to generate public databases of chemical information[37], and the establishment by the US Government of Molecular Libraries Screening Centers. The latter initiative is designed to give public-sector researchers access to an initial library of around 500,000 small molecules for use in probing a diverse range of biological systems. These compounds may lead to new research tools and could aid the development of new drugs or the discovery of new applications for existing ones (see NIH Molecular Libraries Initiative, http://nihroadmap.nih.gov).

To exploit fully the emerging chemical tools and new methodologies in molecular and structural biology (for example, http://www.nigms.nih.gov/psi/centers.html), and so make the quantum leap in the efficiency of drug discovery that these developments promise, chemists must increasingly develop strong interactions with scientists from different disciplines. With such interdisciplinary collaborations it will be possible to embrace some of the grand challenges that exist in our quest to understand and manipulate the chemistry of life for the benefit of mankind. One of the greatest challenges must be to discover and understand what fraction of the universe of chemical space is used by living systems, and how much more could in principle be used to influence these systems. Progress in this area of science will lead to more efficient strategies for drug discovery. And as such challenges are embraced, we shall very likely learn many of the secrets of how life began and evolved. □

doi:10.1038/nature03192

1. Fersht, A. R. *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding* (W. H. Freeman, New York, 1999).
2. Bohacek, R. S., McMartin, C. & Guida, W. C. The art and practice of structure-based drug design: a molecular modelling perspective. *Med. Res. Rev.* **16,** 3–50 (1996).
3. Luria, S. E., Gould, S. J. & Singer, S. *A View of Life* (Benjamin/Cummings, Menlo Park, California, 1981).
4. Goto, S., Okuno, Y., Hattori, M., Nishioka, T. & Kanehisa, M. LIGAND: database of chemical compounds and reactions in biological pathways. *Nucleic Acids Res.* **30,** 402–404 (2002).
5. Lander E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 806–921 (2001).
6. Ellis R. J. & Minton, A. P. Join the crowd. *Nature* **425,** 27–28 (2003).
7. Hall, D. & Minton, A. P. Macromolecular crowding: qualitative and semiquantitative successes, quantitative challenges. *Biochim. Biophys. Acta.* **1649,** 127–139 (2003).
8. Voet, D. & Voet, J. G. *Biochemistry* 2nd edn (Wiley, New York, 1995).
9. Goodsell, D. S. Inside a living cell. *Trends Biochem. Sci.* **16,** 203–206 (1991).
10. Westerhoff, H. V. & Palsson, B. O. The evolution of molecular biology into systems biology. *Nature Biotechnol.* **22,** 1249–1252 (2004).
11. Skår, J. & Coveney, P. V. Self-organization: the quest for the origin and evolution of structure. *Phil. Trans. R. Soc. Lond. A* **361,** 1047–1317 (2003).
12. Tsien, R. Y. Imagining imaging's future. *Nature Rev. Mol. Cell Biol.* **4,** SS16–SS21 (2003).
13. Medalia, O. *et al.* Macromolecular architecture in eukaryotic cells visualised by cryoelectron tomography. *Science* **298,** 1209–1213 (2002).
14. Kitano, H. Computational systems biology. *Nature* **420,** 206–210 (2001).
15. Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* **23,** 3–25 (1997).
16. Houghten, R. A. Parallel array and mixture-based synthetic combinatorial chemistry: tools for the next millenium. *Annu. Rev. Pharmacol. Toxicol.* **40,** 273–282 (2000).
17. Schreiber, S. L. Target-oriented and diversity-oriented organic synthesis in drug discovery. *Science* **287,** 1964–1969 (2000).
18. Bleicher, K. H. *et al.* Hit and lead generation: beyond high-throughput screening. *Nature Rev. Drug Discov.* **2,** 369–378 (2003).
19. Feher, M. & Schmidt, J. M. Property distributions: differences between drugs, natural products and molecules from combinatorial chemistry. *J. Chem. Inf. Comput. Sci.* **43,** 218–227 (2003).
20. Service, R. F. Surviving the blockbuster syndrome. *Science* **303,** 1796–1799 (2004).
21. Dickson, M. & Gagnon, J. P. Key factors in the rising cost of new drug discovery and development. *Nature Rev. Drug Discov.* **3,** 417–429 (2004).
22. Mullin, R. Drug Discovery. *Chem. Eng. News* **82,** 23–31 (2004).
23. Collins, F. S. *et al.* A vision for the future of genomics research. *Nature* **422,** 835–847 (2003).
24. Zambrowicz, B. P. & Sands, A. T. Knockouts model the 100 best-selling drugs — will they model the next 100? *Nature Rev. Drug Discov.* **2,** 38–51 (2003).
25. Hopkins, A. L. & Groom, C. R. The druggable genome. *Nature Rev. Drug Discov.* **1,** 737–730 (2002).
26. Drews, J. Drug discovery: a historical perspective. *Science* **287,** 1960–1964 (2000).
27. Newman, D. J., Cragg, G. M. & Snader, K. M. Natural products as a source of new drugs over the period 1981–2002. *J. Nat. Prod.* **66,** 1002–1037 (2002).
28. Pace, N. R. A molecular view of microbial diversity and the biosphere. *Science* **276,** 734–740 (1997).
29. Schreiber, S. L. Chemical genetics resulting from a passion for synthetic organic chemistry. *Bioorg. Med. Chem.* 6, 1127–1153 (1998).
30. Sills, M. A. *et al.* Comparison of assay technologies for a tyrosine kinase assay generates different results in high throughput screening. *J. Biomol. Screening* **7,** 191–214 (2002).
31. Steitz, T. A. & Moore, P. B. RNA, the first macromolecular catalyst: the ribosome is a ribozyme. *Trends Biochem. Sci.* **28,** 411–418 (2003).
32. Novina C. D. & Sharp, P. A. The RNAi revolution. *Nature* **430,** 161–164 (2004).
33. Seeberger, P. H. Automated carbohydrate synthesis to drive chemical glycomics. *Chem. Commun.* **10,** 1115–1121 (2003).
34. Buckingham, S. Bioinformatics: programmed for success. *Nature* **425,** 209–215 (2003).
35. Agrafiotis, D. K., Lobanov, V. S. & Salemme, F. R. Combinatorial informatics in the post-genomics era. *Nature Rev. Drug Discov.* **1,** 337–346 (2002).
36. Townsend J. A. *et al.* Chemical documents: machine understanding and automated information extraction. *Org. Biomol. Chem.* **22,** 294–300 (2004).
37. Schreiber, S. L. The small-molecule approach to biology: chemical genetics and diversity-oriented organic synthesis make possible the systematic exploration of biology. *Chem. Eng. News* **81,** 51–61 (2003).
38. Richardson, J. S. & Richardson, D. C. Natural beta-sheet proteins use negative design to avoid edge-to-edge aggregation. *Proc. Natl Acad. Sci. USA* **99,** 2754–2759 (2002).
39. Dobson, C. M., Gerrard J. A. & Pratt, A. J. *Foundations of Chemical Biology* (Oxford Univ. Press, Oxford, 2001).
40. *Nature* insight: protein misfolding *Nature* **426,** 883–909 (2003).
41. Dobson C. M. Protein folding and misfolding. *Nature* **426,** 884–890 (2003).
42. Williams, R. J. P. & Frausto da Silva, J. J. R. *The Natural Selection of the Chemical Elements* (Oxford Univ. Press, Oxford, 1997).
43. Westheimer, F. H. Why nature chose phosphates. *Science* **235,** 1173–1178 (1987).
44. Chin J. W. *et al.* An expanded eukaryotic genetic code. *Science* **301,** 964–967 (2003).

Acknowledgements I thank the Wellcome and Leverhulme Trusts for their support through programme grants.

Competing interests statement The author declares that he has no competing financial interests.